AusTalk: an audio-visual corpus of Australian English

Dominique Estival[1], Steve Cassidy[2], Felicity Cox[2], Denis Burnham[1]
[1] MARCS Institute, U. of Western Sydney, Australia
[2] Macquarie University Australia
E-mail: d.estival@uws.edu.au, steve.cassidy@mq.edu.au, felicity.cox@mq.edu.au, d.burnham@uws.edu.au

Abstract

This paper describes the *AusTalk* corpus, which was designed and created through the *Big ASC*, a collaborative project with the two main goals of providing a standardised infrastructure for audio-visual recordings in Australia and of producing a large audio-visual corpus of Australian English, with 3 hours of AV recordings for 1000 speakers. We first present the overall project, then describe the corpus itself and its components, the strict data collection protocol with high levels of standardisation and automation, and the processes put in place for quality control. We also discuss the annotation phase of the project, along with its goals and challenges; a major contribution of the project has been to explore procedures for automating annotations and we present our solutions. We conclude with the current status of the corpus and with some examples of research already conducted with this new resource. *AusTalk* is one of the corpora included in the *Alveo* Virtual Lab, which is briefly sketched in the conclusion.

Keywords: audio-visual corpus, Australian English, standardised infrastructure, collection protocol, annotations.

1. The Big ASC project

The 'Big Australian Speech Corpus (Big ASC)' is a collaborative project between 11 institutions, funded by an Australian Research Council Linkage Infrastructure, Equipment and Facilities grant (Burnham, Cassidy, et al. 2009); (Burnham, Ambikairajah, et al. 2009) with the twin goals of 1) providing a standardized infrastructure for audio-visual (AV) recordings and 2) producing a large AV corpus of Australian English (AusE). Up to 1000 geographically and socially diverse speakers are recorded in locations across Australia using 12 sets of standardised hardware and software 'Black Boxes' with a uniform and automated protocol (the Standard Speech Collection Protocol – SSCP) to produce the *AusTalk* corpus (Wagner et al. 2010); (Burnham et al. 2011).

While the main purpose of *AusTal*k is to provide an extensible database for projects charting the extent, degree, and details of social, regional and ethno-cultural variations in AusE, it was also designed to support a range of research projects and applications. In order for the *AusTalk* corpus to cater to the needs of different researchers in various disciplines such as phonetics, forensic studies, speech and language technologies, linguistic analysis, audio-visual analysis, the Big ASC project had to strike a balance between high quality studio recording and field data collection. This was achieved through the strict data collection protocol, with high levels of standardisation and automation, and a recruitment process that ensured sufficient variability.

2. The *AusTalk* Corpus

When complete, the AusTalk corpus will comprise nearly 3000 hours of audio and video recordings from 1000 AusE speakers, all having completed primary and secondary education in Australia (but not necessarily having been born in Australia), a criterion that ensures inclusion of a range of speakers from various cultural backgrounds. According to the 2011 census, 26% of the Australian population were born overseas and a further 20% had at least one parent born overseas (ABS 2012). The criterion that participants had attended school in Australia was to ensure that we only captured speakers of AusE rather than foreign accented English. Demographic and language background information was collected from all participants via an online questionnaire to establish their gender, age, residential and educational history, cultural heritage, hobbies, occupation, languages spoken, parents' details and any speech and hearing difficulties they may have had. Almost 800 speakers have now been recorded at 13 different sites, with more than 2000 sessions uploaded, a total of 22TB of data. Data collection is still proceeding at three sites and will be completed by the end of 2014.

2.1 The *AusTalk* corpus components

Audio-visual corpora are important for different types of research in linguistics, Natural Language Processing (NLP) and Language and Speech Technologies, relying of various aspects of variability. In the *AusTalk* corpus, three one-hour sessions are recorded at intervals of at least one week to capture potential *variability over time*, while *geographical variability* is guaranteed by recording at locations covering all the capital cities of Australian states and territories and several regional centres. Stratified sampling across gender and three broad adult age groups captures *individual variability*. Wide advertising and high visibility of the project, with a well-publicised launch on Australia Day 2011 and good media coverage, helped recruit speakers from a range of social spheres to ensure *social variability*.

The *AusTalk* corpus contains a variety of speech content from a range of tasks, with four Read Speech and five Spontaneous Speech components, and all the data is captured by five microphones and two stereo cameras recording audio and video. Each of the three recording sessions comprises a different subset of the Read and Spontaneous speech tasks.

In the standard 'Words', 'Digits' and 'Sentences' tasks, the speaker reads aloud a list of prompts from a computer screen, while the 'Story Reading' and 'Story Re-telling' tasks (Session 1) provide material for the study

of differences between reading and spontaneous language. The 'Interview', 'Map Task' and 'Conversation' tasks provide material for the analysis of speech acts in dialogues. In the 'Interview' (Session 2), the speakers talk to the Recording Assistants (RAs) on a topic which they chose in Session 1. The 'Map Task' (Session 3) is designed along the lines of (Anderson et al. 1991) but adapted for AusE to contain locations and landmarks selected to sample a range of AusE phonological features. In this third session, two speakers are paired for two Map Tasks, so that each participant plays the role of Information Giver and Information Receiver, after which they discuss the experience in the 'Conversation'. At the beginning and end of each session, a set of natural 'Yes/No' questions elicit a range of positive and negative answers.

Table 1 shows the distribution of these tasks across the sessions and the average time for each task.

Components	Session	Time (mins)	Time per speaker
Read speech			**53 mins**
Words (322 x 3)	S1, S2, S3	10	30
Digit strings (12 x 2)	S1, S2	5	10
Sentences (59 x 1)	S2	8	8
Read story	S1	5	5
Spontaneous speech			**80 mins**
Yes/No answers (x 5)	S1, S2, S3	2	10
Re-told story	S1	10	10
Interview	S2	15	15
Map Task (x 2)	S3	20	40
Conversation	S3	5	5
TOTAL (average)			**133 mins**

Table 1: AusTalk Corpus Components / time per speaker

2.2 Quality Control

To ensure high data quality as well as consistency across all the sites, several processes were put in place. First, before the data collection began, all the Recording Assistants (one for each of the 16 locations) were trained together during a 2-day centrally-located workshop at the University of Western Sydney (UWS), at which they practiced setting up the equipment and running through the recording sessions with each other. Additional training was required when new RAs were recruited, an important factor in maintaining consistency in the data collection.

Second, each recording site initially made sample recordings which were centrally checked for audio and video quality before the start of data collection at that particular site, and modifications were made to the location, to remove sources of noise or modify light conditions if the quality was sub-standard.

Third, there was continuous monitoring of data quality and feedback and advice to the RAs throughout the corpus collection. A Quality Control RA (QC-RA) was employed at the central UWS receiving site where the data was uploaded, with strict guidelines for both audio and video quality checks. To help the site RAs and the QC-RA, we developed the SSCP-QC, a utility to check the number of files along with the quality of parameters such as silence or loudness for audio, and frame skipping or brightness for video. The outcomes of the QC checks are retained and have become part of the published metadata indicating the QC status at the item and component levels. Manual inspection of the data finalises the published QC status, as one of the following:

- A (A-OK)
- B (OK, but imperfect)
- C (bad, not acceptable)
- D (deficient or missing, e.g. "`Missing 2nd video camera for Map Task`")

3. The AusTalk Annotation Task

Two important usability goals of the Big ASC project are to make *AusTalk* widely available and to allow future contributions, such as addition of further data or additional annotations. Audio and video data are stored on a web-accessible server, with corpus metadata and annotations stored in the DADA annotation store (Cassidy and Johnston 2009). The DADA server allows import/export of annotation data in formats supported by many annotation and analysis tools.

The Annotation Task itself could not be commenced until sufficient data were collected and organised (in late April 2012), but is now well under way. In this section, we first delimit the scope of the Annotation Task, then describe the processes we have put in place and the annotations that have already been produced before briefly mentioning the main challenges we faced.

The original goal of the Big ASC project was to annotate all the data collected and, from the beginning of the project, it was decided to automate the annotation process as much as possible, while providing high-quality manual annotation for a subset of the data. It was expected that forced alignment would be used where appropriate to enhance manual annotation. Thus, the Annotation Task

was limited to 1) word segmentation for the Read Speech and 2) orthographic transcription aligned at the phrase or sentence level for Spontaneous Speech. Integral to the project is that new annotations, e.g. detailed phonetic transcriptions or Part-of-Speech tagging, can later be contributed by project partners or other researchers and then integrated into the existing annotation store. We thus defined the goal for annotations as follows:

a) **Orthographic** level annotation of both Read and Spontaneous speech.
   For Read Speech, the script provides the basis for an orthographic transcription.
   For Spontaneous Speech, we were optimistic that orthographic transcripts could be generated from an automatic speech recognition package such a Dragon (DNS).
b) **Phonemic** level transcription with segmentation, to be automated as much as possible.
   For Read Speech, forced alignment would be performed in collaboration with an external partner (Schiel, Draxler, and Harrington 2011), following manual phonemic level transcription of a subset of the data. This subset would provide the training set for the forced aligner.
c) **Audio-Video** alignment. Automatic alignment is provided by the strobe signal recorded on a separate audio channel (Lichtenauer et al. 2009).

The following is the 'wish-list' of annotations to be performed if there were sufficient resources:

d) **Phonetic** level: manual and labour intensive
e) **Intonation**
f) **Part-of-Speech** – to be automated
g) **Morphemic**
h) **Syntactic**

We decided early on that the Big ASC project could only afford to manually annotate a subset of the data: for 5 speakers a full set of read speech data would be annotated (levels a and b above), while 100 additional speakers would only have a subset of their data manually annotated.

Manual annotation is labour-intensive and expensive. Therefore a major contribution of the project has been to explore procedures for automating annotations through collaboration with partners to produce 1) alignment for the Read Speech and 2) orthographic transcriptions for the Spontaneous Speech components.

As was expected, automating the time alignment of phonemic transcriptions for the Read Speech data proved very challenging. A major obstacle for this part of the project was that people make mistakes when reading material aloud, so scripted data does not always have the integrity required for automatic processing. The preliminary manual annotation phases of the project was therefore extended to provide sufficient high quality phonemically-transcribed data that would form the essential training materials for the Australian module of the Munich Automatic Segmentation System MAUS (Schiel, Draxler, and Harrington 2011). To this end the 59 read sentences from 100 speakers were orthographically and phonemically transcribed manually. The MAUS utility which was modified for AusE based on this training set is now capable of returning Praat textgrids (Boersma and Weenink 2001) containing phonemically transcribed and segmented data upon presentation of orthographic input. These automatically generated Praat textgrids can be manually corrected where necessary and provide a platform for more detailed segmentation and labelling in the future.

It proved too great a challenge to generate automatic orthographic transcriptions for Spontaneous Speech, so a third party transcription company was contracted to produce transcripts for the same sample of 100 speakers as above. It is then possible to pass the orthographic transcriptions through MAUS, which will return automatically generated textgrids for the Spontaneous Speech data.

In order to set a standard for annotation quality, the full set of scripted speech data from five speakers has been manually phonemically segmented and labelled in Praat by a team of highly trained annotators. We have created a purpose-built annotation manual to ensure consistency across annotators and tasks. Sentence data from an additional 30 speakers have also been manually annotated. These manually labelled data provide a benchmark that can be used as training material for further manual correction of automatically generated data.

4. Conclusion and Future Work

The data collection phase for *AusTalk* is coming to an end, with less than 200 speakers remaining to be recorded at three sites in order to complete the full complement of three one-hour sessions for 1000 AusE speakers.

Table 2 shows the distribution of speakers across recording sites. The figures given in the "Actual" columns show the number of speakers for whom data has been not only recorded but uploaded to the server. Detailed demographic statistics concerning gender, age and education level of the speakers are made available on the data server.

STATE	Capital Cities (University)	Target	Actual	Regional Centres (University)	Target	Actual	Other	Target	Actual
NSW	Sydney (USYD)	64	**64**	Armidale (UNE)	48	**44**	Emotions (UNSW)	36	0
	Sydney (UNSW)	48	**48**	Bathurst (CSU)	48	**46**			
QLD	Brisbane (UQ)	100	**75**	Townsville (UQ)	48	**30**			
				Maroochydore (USC)	20	**20**			
VIC	Melbourne (UMELB)	120	**118**	Castlemaine (UMELB)	48	**22**			
SA	Adelaide (Flinders)	96	**96**						
NT				Darwin (CDU)	24	**2**	Aboriginal English (CDU)	48	0
				Alice Springs (CDU)	24	**0**			
WA	Perth (UWA)	96	**96**						
TAS	Hobart (UTAS)	48	**48**						
ACT	Canberra (UC)	36	**39**						
	Canberra (ANU)	48	**48**						
Totals		656	**632**		260	**164**		84	0

Table 2. Distribution of recorded and uploaded speakers data (March 2014)

Follow-on projects have already begun to collect data from different population groups in some locations (e.g. particular ethnic backgrounds in Canberra) and the analysis of *AusTalk* data is under way at other partner sites, e.g. video analysis for facial gestures (Sui et al. 2012a, 2012b) and close phonetic analysis of the isolated word list data.

There will be a tutorial at Interspeech 2014 (Togneri, Bennamoun, and Sui 2014) in which attendees will learn how to use the 3D based AV corpus derived from *AusTalk* for audio-visual speech/speaker recognition. Experimental results using this corpus show that, compared with the conventional AVSR based on the audio and grey-level visual features, there is a significant speech accuracy increase by integrating both depth-level and grey-level visual features.

In a study based on the framework of (Weiss, Burkhardt, and Geier 2013), the Read Sentences provide a rich body of stimuli used to study perceptual dimensions used by listeners to characterise speakers' vocal characteristics and speaking style. By collecting similarity measures for triplets of stimuli from 13 male speakers with an incomplete design (Burton and Nerlove 1976), fundamental perceptional dimensions separating these speakers can be extracted by applying multi-dimensional scaling on perception data from 15 male non-experts listeners. Along with the similarity decisions, individual labels for each triplet are assessed as a starting point for interpreting the dimensions found, as well as providing additional material to develop a questionnaire describing speakers' vocal characteristics and speaking style.

An unforeseen but very exciting addition to *AusTalk* was the inclusion of speakers who originally participated in the Australian National Database of Spoken Language (ANDOSL) project in 1993-95 (Vonwiller et al. 1995). Of the eight ANDOSL speakers who were found and who agreed to participate in AusTalk, four completed the full AusTalk recording sessions and these constitute invaluable longitudinal data for the study of AusE.

Annotation and quality assessment continue as more data are collected and made available through a new interface. Annotation is an important aspect of the Big ASC and other similar projects for, without it, many of the applications such as Automatic Speech Recognition, and much of the proposed research could not be conducted. While the ideal of providing full annotations of 100% of the data will not be realised in this phase of the project, we are providing a full set of manually created phonemic and orthographic transcriptions for a selected number of speakers. We will also provide automatically time-aligned transcriptions for all the Read Speech data and automatically generated orthographic transcriptions for at least a subset of the Spontaneous Speech data. Together these will constitute the basis and a protocol for further annotation in the future.

Meanwhile, the *AusTalk* corpus is already included in *Alveo,* the Human Communication Science Virtual Laboratory, a recent NeCTAR-funded Australian collaborative project that aims to provide a platform for easy access to language, speech and other communication-relevant databases and for the integrated use of a range of analysis tools (Burnham et al. 2012). *Alveo* incorporates existing tools, some developed by project members, which were adapted to work on the shared infrastructure, together with a web-based data discovery interface for searching and accessing the text,

speech, AV and music datasets contributed by the project partners. The tools are orchestrated by a workflow engine with both web and command line interfaces to allow use by technical and non-technical researchers (Cassidy et al. 2014). *Alveo* will allow the generation of automated Part-of-Speech tagging and syntactic analyses as additional annotations for the *AusTalk* corpus.

5. Acknowledgements

We gratefully acknowledge financial and/or in-kind assistance of the Australian Research Council (LE100100211), ASSTA; the Universities of Western Sydney, Canberra, Melbourne, NSW, Queensland, Sydney, Tasmania and Western Australia; Macquarie, Australian National, and Flinders Universities; and the Max Planck Institute for Psycholinguistics, Nijmegen.

6. References

ABS. 2012. Cultural Diversity in Australia. http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2071.0main+features902012-2013. Australian Bureau of Statistics.

Anderson, A.H., M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, and R. Weinert. 1991. "The HCRC Map Task Corpus." *Language and Speech* no. 34 (4):351-366.

Boersma, Paul, and David Weenink. 2001. "Praat, a system for doing phonetics by computer." *Glot International* no. 5 (9/10):341-345.

Burnham, Denis, E. Ambikairajah, Joanne. Arciuli, Mohammed Bennamoun, C.T. Best, Steven Bird, A.B. Butcher, Steve Cassidy, G. Chetty, F.M. Cox, Anne Cutler, Robert Dale, Julien R. Epps, Janet M. Fletcher, Roland Goecke, David B. Grayden, John T. Hajek, John C. Ingram, Shun Ishihara, Nenagh Kemp, Yuko Kinoshita, T. Kuratate, T.W. Lewis, D.E. Loakes, Mark Onslow, David M. Powers, P. Rose, Roberto Togneri, D. Tran, and Michael Wagner. 2009. A blueprint for a comprehensive Australian English auditory-visual speech corpus. In *2008 HCSNet Workshop on Designing the Australian National Corpus*. Sydney: Somerville, MA, USA: Cascadilla Proceedings Project.

Burnham, Denis, Steve Cassidy, Felicity Cox, and Robert Dale. 2009. The Big Australian Speech Corpus: An Audio-Visual Speech Corpus of Australian English Australian Re-search Council Linkage, Infrastructure, Equipment and Facilities Grant. Original edition, LE100100211.

Burnham, Denis, Dominique Estival, Peter Bugeia, Peter Sefton, and Steven Cassidy. 2012. Above and Beyond Speech, Language and Music: A Virtual Lab for Human Communication Science (HCS vLab). NeCTAR (National eResearch Collaboration Tools & Resources) Virtual Laboratory. Original edition, VL222.

Burnham, Denis, Dominique Estival, Steven Fazio, Felicity Cox, Robert Dale, Jette Viethen, Steve Cassidy, Julien Epps, Roberto Togneri, Yuko Kinoshita, Roland Göcke, Joanne Arciuli, Marc Onslow, Trent Lewis, Andy Butcher, John Hajek, and Michael Wagner. 2011. Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box. In *Interspeech 2011*. Florence, Italy.

Cassidy, Steve, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. 2014. The Human Communication Science Virtual Laboratory: A Web Based Repository API. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland.

Cassidy, Steve, and Trevor Johnston. 2009. Ingesting the Auslan Corpus into the DADA Annotation Store. In *Third Linguistic Annotation Workshop (LAW III)*. Singapore.

Lichtenauer, Jeroen, Michel Valstar, Jie Shen, and Maja Pantic. 2009. Cost-Effective Solution to Synchronized Audio-Visual Capture Using Multiple Sensors. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09)*. Washington, DC, USA: IEEE Computer Society.

Schiel, Florian, Christoph Draxler, and Jonathan Harrington. 2011. Phonemic Segmentation and Labelling using the MAUS Technique. In *Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research'*. University of Pennsylvania, Philadelphia PA. USA.

Sui, Chao, Serajul Haque, Roberto Togneri, and Mohammed Bennamoun. 2012a. A 3D Audio-Visual Corpus for Speech Recognition. In *SST2012*. Sydney, Australia: ASSTA.

Sui, Chao, Serajul Haque, Roberto Togneri, and Mohammed Bennamoun. 2012b. Discrimination Comparison Between Audio and Visual Features. In *Asilomar 2012*. Pacific Grove, USA.

Togneri, Roberto, Mohammed Bennamoun, and Chao Sui. 2014. Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus. Tutorial at Interspeech 2014. Singapore.

Vonwiller, J., I. Rogers, C. Cleirigh, and W. Lewis. 1995. "Speaker and Material Selection for the Australian National Database of Spoken Language." *Journal of Quantitative Linguistics* no. 3:177-211.

Wagner, M., D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, D. Loakes, T. Lewis, T. Kuratate, Y. Kinoshita, N. Kemp, S. Ishihara, J. Ingram, J. Hajek, D.B. Grayden, R. Göcke, J. Fletcher, D. Estival, J. Epps, R. Dale, A. Cutler, F. Cox, G. Chetty, S. Cassidy, A. Butcher, D. Burnham, S. Bird, C. Best, M. Bennamoun, J. Arciuli, and E. Ambikairajah. 2010. The Big Australian Speech Corpus (The Big ASC). In *13th Australasian International Conference on Speech Science and Technology*, edited by M. Tabain, J. Fletcher, D. Grayden, Hajek J. and A. Butcher. Melbourne: ASSTA.

LREC 2014, Ninth International Conference on Language Resources and Evaluation

United Nations Educational, Scientific and Cultural Organization Under the patronage of **UNESCO**	**May 26-31, 2014** Harpa Concert Hall and Conference Center Reykjavik, Iceland
Editors: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. **Assistant Editors:** Hélène Mazo, Sara Goggi	 ISBN 978-2-9517408-8-4 EAN 9782951740884

Organization:

European Language Resources Association

Evaluation and Language resources Distribution Agency

Istituto di Linguistica Computazionale

LREC 2014 Committees

LREC 2014 Programme at a Glance

Archive Content:

This archive contains the 742 papers accepted for presentation at LREC 2014. All papers are in PDF format. Please read instructions below to see how you can read the files.

You can use the navigation menu located on the top of this page to browse for papers in various ways:

"**Sessions**" contains introductory messages, keynote speeches, panels information, and titles of papers organized by session.

"**Papers**" contains a list of the papers organized by their title.

"**Authors**" contains a list of all authors and their related papers.

"**Workshops**" contains a list of all workshops and tutorials.

"**Topics**" contains a list of all topics and their related papers.

"**Affiliations**" contains a list of all affiliations and their related papers.

Sponsors:

Silver Sponsors	Bronze Sponsors	
Holmes semantic solutions	VoiceBox TECHNOLOGIES	European Media Laboratory GmbH EML
	KDICTIONARIES	IMI

Support:

Some of the content on this CD requires JavaScript to be enabled in your web browser to work properly. This includes the menu navigation and the search engine, but is not limited to. To enjoy the full interactivity of the CD, please ensure that JavaScript is enabled in your browser.

Installation files for Abode Acrobat Reader are included in this CD for your convenience :

Acrobat Reader for Windows users (English language)
Acrobat Reader for Linux users (English language)
Acrobat Reader for MacOS Intel users (English language)
Acrobat Reader for MacOS PowerPC users (English language)

If you want Adobe Acrobat Reader in your language or for other platfoms, you can download the latest version directly on Adobe website: http://www.adobe.com/

Introduction of the Conference Chair and ELRA President
Nicoletta Calzolari

I wish to express to Mrs. Irina Bokova, Director-General of UNESCO, the gratitude of the Program Committee, of all LREC participants and my personal for her Distinguished Patronage of LREC 2014. Languages – mentioned in the first article of UNESCO Constitution – have been at the heart of UNESCO mission and programmes throughout its history.

I am also especially grateful to Madame Vigdís Finnbogadóttir, UNESCO's Goodwill Ambassador for languages and former President of Iceland (1980-1996), first woman in the world elected as head of state in a democratic election, for the continuous personal support she has granted to LREC since our first visit in Reykjavík in 2012. In her name the Vigdís International Centre for Multilingualism and Intercultural Understanding has been established under the auspices of UNESCO to promote multilingualism and raise awareness of the importance of language as a core element of the cultural heritage of humanity. I quote a sentence from a recent interview where she says: "The land—our nature—and language, those are our national treasures": this tells a lot of why this LREC is in Iceland!

Some figures: all records broken!

LREC 2014, the 9th LREC, with its 1227 submissions, has set a new record! We received 21% more submissions than in 2012. We continue the tradition of breaking our own previous records: out of the 1227 submissions, after the reviewing process by well 970 colleagues, we accepted 745 papers. We also accepted 22 workshops and 9 tutorials. More than 1100 participants have already registered at the beginning of May.

These figures have a meaning. The field of Language Resources and Evaluation is continuously growing. And LREC continues to be – as many say – "the conference where you have to be and where you meet everyone".

Every time I underline the fact that a relatively high acceptance rate (60.7% this time) is for us a reasoned choice. It is important to get a pulse on the situation, to monitor the evolution of the field in the many varieties of approaches and methodologies, and in particular for many different languages. For us, a lexicon in any language is as important as a lexicon in American English. Multilingualism – and equal treatment of all languages – is a feature at the heart of LREC. Other venues promote a sense of exclusivity (also through the equation low acceptance rate and great merit); we always encourage a sense of inclusiveness. This is a typical feature of LREC that makes it a special conference. Quality is not necessarily undermined by a high acceptance rate, but also by the influence of the papers on the community: the ranking of LREC among other conferences in the same area proves this. According to Google Scholar h-index, LREC ranks 4th in Computational Linguistics at a similar level of conferences using much lower acceptance rates, just like the LRE Journal also ranks 4th in the general field of Humanities, Literature and Arts.

LREC 2014 Trends

Language Resources (LRs) being everywhere in Language Technology (LT), LREC is a perfect observation point of the evolution of the field. Looking at all the topics, while building the program and putting all the pieces together, the most striking (even if not surprising) new trend was for me the application of sentiment/opinion discovery/analysis to social media shown by so many papers.

A very rough sketch of LREC 2014 major topics and trends, from my viewpoint, is the following:

- There is a completely new topic:
- Linked Data, also the hot topic of this edition
- Topics that were quite new in 2012 and are now consolidated:
- Social Media, in particular combined with subjectivity, as said above
- Crowdsourcing and Collaborative Construction of LRs
- Other increasing (not the biggest in absolute terms) topics with respect to last LREC are:
- Subjectivity: Sentiments, Emotions, Opinions
- Less-resourced languages, in line with the value we give to safeguarding world's linguistic diversity
- Extraction of Information, Knowledge discovery, Text mining: always a very hot topic
- Computer Aided Language Learning
- Stable Big topics:
- Infrastructural issues and Large projects, and also Standards and Metadata, receive the usual attention by the LREC authors
- Lexicons and Corpora (i.e. the most typical "data"), of many types, modalities and for many purposes and applications: they are the prominent and most crowded topic
- Semantics and Knowledge, in all their variations: from annotation of anaphoric information, to ontologies and WordNets, sense disambiguation, named entities recognition, information extraction, to mention just a few
- Syntax, Grammar and Parsing continues to be a largely represented topic: not solved
- Machine Translation and Multilingualism are areas on which a lot of work is carried out
- Speech and Multimodality keep the same level: good but not enough
- Dialogue and discourse, with contributions from both the Speech and Text communities
- Evaluation is pervasive/everywhere: we are proud to give evidence to its being an essential feature in the LT landscape
- Tools, systems for text analysis and applications are presented in many papers

A usual observation is the relevance of *infrastructural issues* and the attention that LREC – and ELRA – pay to them. They are mostly neglected in other conferences. Infrastructural issues play an important role for the field of LRs and for the LT field at large. But it is a fact that the first to recognise their importance have been people of the LR area. LRs are themselves of infrastructural nature and quite naturally call for attention to these issues. The infrastructural nature of LRs, captured by the term "Resources", was highlighted in the Introduction of Antonio Zampolli to the 1st LREC in Granada in 1998.

The fact that so many topics are represented at LREC means also that all the various LR and LT sub-communities are present at LREC: this increases the LREC impact and gives to LREC the characteristic of being a true melting pot of cultures, and an enabler of new cooperation initiatives.

15th LREC Anniversary

LREC was born in 1998 and on the occasion of its 15th Anniversary, Joseph Mariani has prepared an analysis of all the past LREC Proceedings, rediscovering the dynamism of the field while looking at the major contributors, topics, trends, also comparing them with an analogous survey done for the speech community on the Interspeech conference series. There will also be a Quiz for all the LREC participants and a winner! The survey paper is in the Proceedings as a special paper for the 15th Anniversary and will be presented at the Closing Session.

ELRA and LREC: a tradition of innovations at the service of our community

I am proud to announce a number of recent initiatives of ELRA and LREC that touch topics that are at the forefront of a paradigm shift and together help advance our field and increase confidence in scientific results. As an introduction I use some words of Zampolli in 1998: "The need to preserve, actively promote the use of, and effectively distribute LR, has caused the USA and EU authorities to put in place, respectively, LDC (the Linguistic Data Consortium) and ELRA (the European Language Resources Association)", observing also that their activities "demand regular updating to reflect technical and strategical evolution of their environment". We try to keep with this recommendation.

These innovations – introduced by ELRA and /or LREC – must not be seen as unrelated steps, but as part of a coherent vision, promoting a new culture in our community. We want to encourage also in the field of LT and LRs what is in use in more mature sciences and ensure reproducibility as a normal part of scientific practice. We try thus to influence how our science is organised or should be organised in the future.

I give here a quick picture of some innovations that are critical for the research process and constitute a sort of manifesto for a new kind of sustainability plan around LRs.

LRE Map

The LRE Map (http://www.resourcebook.eu/), started in 2010, is now an established tool, consulted every day and used in other major conferences. At this LREC we have collected by the authors descriptions for more than 1000 resources in more than 150 languages!

Spreading the LR documentation effort across many people, instead of leaving it only in the hands of the LR distribution centres, we also encourage awareness of the importance of metadata and proper documentation. Documenting a LR is the first step towards identifiability, which in its turn is the first step towards reproducibility.

Recognising the value of Linked Data, we just published the LRE Map in LOD (Linked Open Data).

Share your Language Resources and Reproducibility of research results: the vision

After encouraging sharing LR metadata, the next step is sharing the actual content. ELRA has embraced in the last years the notion of "open LRs": we show this also with the "Share your Language Resources" initiative started in this LREC. With it we ask all the authors to consider making background data available with their paper. More than 300 LRs have been made available: a big success for the first experiment! Showing the community commitment to sharing.

LRE Map and Share your LRs must be seen not as isolated initiatives, but as complementary steps towards implementing a new vision of the field. On one side we encourage opening data that could be valuable to others, on the other we try to encourage a sort of cultural change in our community.

Here the *vision*: It must become common practice also in our field that in conferences and journals when you submit a paper you are offered the opportunity to upload the LRs related to your paper. We must unlock the material that lies behind the papers: the adoption of such a policy will make the whole picture clearer. We had to fight in the '90s for concepts like "reusability", which finally led to promoting the need of developing standards in our field (this was still a hot topic in 1998 at the time of the 1st LREC). Now the need for standards is consolidated and we consider it normal, but we need to start another campaign for encouraging more resource sharing. Researchers are not yet sharing very well; they tend to hold back knowledge. I hope that this sharing trend will be more easily embraced by younger colleagues who are familiar with everyday use of social media of all sorts and free ideas sharing: we must port the same attitude in the research environment. This will fundamentally change the way of making science, in a sort of light revolution towards openness of science in all its facets. Hopefully it will diminish the unfortunate phenomenon of reinventing the wheel from time to time, instead of building on your colleagues' findings.

This vision has to do with many important aspects: shifting to a culture of sharing, re-use, reproducibility of research results. If we want to become a mature science we should make data sharing become "normal" practice. Even more important in a data-intensive discipline like LT. The small cost that each of us will pay to document, share, etc. should be paid back benefiting of others' efforts and become worthwhile. This will also lead to a greater opportunity of collaboration, encouraging bigger experiments by larger collaborative teams (something else we should learn from more mature sciences). Moreover, reproducibility encourages trust.

ISLRN

A major achievement of ELRA has been the recent establishment of the *International Standard Language Resource Number* (ISLRN) (http://www.elra.info/Establishing-the-ISLRN.html). It is a unique identifier to be assigned to each LR. Organised and sustained by ELRA, LDC and AFNLP/Oriental-COCOSDA, the ISLRN Portal provides unique identifiers to LRs. LRs in the ELRA

and LDC catalogues have been the first to get an ISLRN (just one if a LR is stored in both catalogues!).

When you publish a LR it can get an ISLRN and thus become a citable product of research.
Data/LR citation must become normal scientific practice also in our field, as it is in others. To make a LR citable can then pave the way to the design of a sort of "impact factor" of LRs. This can become an important incentive for the field, so that researchers can get the credit they deserve also for the LRs they developed.

ISLRN is not only linked to the possibility of getting proper "recognition" for LR developers. It would also enhance experiment replicability, an essential feature of scientific work. It may thus become a very important advance in our field.

META-SHARE sustainability by ELRA
Through these initiatives we try to encourage community efforts towards: documentation of LRs, possibility of identification of LRs, LR sharing, making research results reproducible. There is a lot of buzz these days around these types of topics. As I said above, all these initiatives are closely related and must become integrated with each other.

For them to become common research practices these activities must be well organised and require good mechanisms behind to become possibly a set of related services on a common platform. Pooling together data from all the research described in conference and journal papers will obviously need an infrastructure for distributing research results and such a LR platform must be sustained.

The ELRA Board has decided to support the META-SHARE platform, but META-SHARE – as sustained by ELRA – must in turn be adapted to be able to support these types of initiatives and thus become also a platform for sharing reproducible research results. We must find ways to make these practices as easy as possible and rewarding for the researcher. META-SHARE – in ELRA view – should become also the obvious repository (recognised by the community) where all these types of actions are sustained and where all research results become available, discoverable, identifiable, and citable. ELRA is taking these steps to start enabling to keep track of connected research activities like papers and supporting underlying resources, in an all-inclusive way.

LREC Proceedings in Thomson Citation Index
A great recent achievement for ELRA and LREC has been the fact that the LREC 2010 and LREC 2012 Proceedings have been accepted for inclusion in CPCI (Thomson Reuters Conference Proceedings Citation Index). This is a significant achievement for LREC and it will provide all LREC authors with a deserved recognition. It is for us of great satisfaction, in particular for the benefit it can bring to young colleagues.

ELRA 18th anniversary and NLP12
Coordination is an important issue when infrastructural issues are at stake. None of the actions above can or should be conducted and tackled in isolation.

For this reason we – ELRA – organised, on the occasion of ELRA majority as its 18th anniversary, the first meeting of the major associations/organisations in the field of Language Resources and Technologies, Computational Linguistics, Spoken Language Processing, Big Data and Digital Humanities, the so-called NLP12 (http://www.elra.info/NLP12-Paris-Declaration.html). We started to discuss issues of common interest to coordinate some of the activities and we adopted some common resolutions, such as the encouragement of language resources and tools sharing and promotion of best practices for language resource citation in publications.

Together we should be able to take the necessary steps to better serve the field and the respective communities and to strengthen the bridges between various communities (e.g. Language Technology and Humanities).

ELRA for Open science

I am excited and proud that we – as ELRA and LREC – can contribute to such a (quiet) revolution towards shaping a new type of *open scientific information space* for the future of our field, the Language Resources and Technology future. I have always felt it is our duty to use the means that we have in our hands to try to shape the future of the field, and in this case to play a role in how to change scientific practice and have an impact on the overall scientific enterprise!

Trying to be always forward-looking and to act in a proactive way to serve the field, ELRA continues to be a community-aware association. I would like to work for it to become more also a community-driven association. We would like to discuss with all those who are interested about how to tackle the challenge of truly open research (which is more than open access!) so that we can take the necessary further steps to make this process more efficient, faster and more collaborative.

It is clear that in such a campaign for the cause of reproducibility and open science and for a proper system of attribution and citation – two closely related aspects– we must involve also funding agencies that should help in supporting the necessary policy actions. For sure we will involve in this initiative the NLP12 group. But I strongly believe that the most important change must come from the mind-set of researchers. This is where LREC can help, I hope …

The message that ELRA has for its community, the LREC community, is: We are here to help!

Acknowledgments

In this last part I wish to express my deepest gratitude to all those who made this LREC 2014 possible and hopefully successful.

I first thank the Program Committee members, not only for their dedication in the huge task of selecting the papers, but also for the constant involvement in the various aspects around LREC. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program. To Joseph Mariani for his always wise suggestions. And obviously to Khalid Choukri, who is in charge of so many aspects around LREC.

I thank ELRA and the ELRA Board: LREC is a major service from ELRA to all the community!

A very special thanks goes to Sara Goggi and Hélène Mazo, the two Chairs of the Organising Committee, for all the work they do with so much dedication and competence, and also the capacity to tackle the many big and small problems of such a large conference (not an easy task). They are the two pillars of LREC, without whose commitment for many months LREC would not happen. So much of LREC organisation is on their shoulders, and it is visible to all participants.

A particular expression of gratitude goes to the Local Committee, and especially to Eiríkur Rögnvaldsson (its Chair) and Sigrún Helgadóttir: they have worked with great commitment and enthusiasm for many months for the success of LREC always looking at the best solutions to the many local issues.

All my appreciation goes also to the distinguished members of the Local Advisory Board for their constant support.

Among the Icelanders I wish to mention Guðrún Magnúsdóttir, for a very simple reason: the idea of having LREC in Iceland came out during a lunch that the two of us had together in Berlin!

I express my gratitude to the Sponsors that believe in the importance of our conference, and have helped with financial support. I am grateful to the authorities, and all associations, organisations, companies that have supported LREC in various ways, for their important cooperation.

Furthermore, on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a wonderful job.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events.

A big thanks goes to all the LREC authors, who provide the "substance" to LREC, and give us such a broad picture of the field.

I finally thank the two institutions that have dedicated such a great effort to this LREC, as to the previous ones, i.e. ELDA in Paris and ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least, thanks are thus, in addition to Hélène Mazo and Sara Goggi, to all the others who have helped and will help during the conference: Victoria Arranz, Paola Baroni, Roberto Bartolini, Irene De Felice, Riccardo Del Gratta, Francesca Frontini, Ioanna Giannopoulou, Johann Gorlier, Olivier Hamon, Jérémy Leixa, Valerie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller. You will meet most of them during the conference.

I also hope that funding agencies will be impressed by the quality and quantity of initiatives in our sector that LREC displays, and by the fact that the field attracts all the best groups of R&D from all continents. The success of LREC for us actually means the success of the field of Language Resources and Evaluation.

And lastly, my final words of appreciation are for all the LREC 2014 participants. Now LREC is in your hands. You are the true protagonist of LREC; we have worked for you all and you will make this LREC great. I hope that you discover new paths, that you perceive the ferment and liveliness of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation … which you will show at the next LREC.

LREC is not exactly in a Mediterranean location this time, even if all the hot water around gives some Mediterranean flavour! But the tradition of holding LREC in wonderful locations continues, and Reykjavík is a perfect LREC location! I am sure you will like Reykjavík and the friendliness of Icelanders. And I hope that Reykjavík will appreciate the invasion of LRECers!
With all the Programme Committee, I welcome you at LREC 2014 in such a wonderful country as Iceland and wish you a fruitful Conference.

Enjoy LREC 2014 in Reykjavík!

Nicoletta Calzolari

Chair of the 9th International Conference on Language Resources and Evaluation and ELRA President